
Article:
Campbell, M.J. orcid.org/0000-0003-3529-2739 (2017) What is propensity score modelling? Emergency Medicine Journal, 34 (3). pp. 129-131. ISSN 1472-0205

https://doi.org/10.1136/emermed-2016-206542

What is propensity score modelling?

Michael J Campbell

Emeritus Professor of Medical Statistics
Design, Trials and Statistics, ScHARR,
University of Sheffield S1 4DA UK
Email: m.j.campbell@sheffield.ac.uk Tel: 0114 230 1837

Keywords: statistics, anaesthesia, research methods

Word count: 2244

Boxes 2

The paper by Caruana et al [1] looks at the effect of cricoid pressure, (used to prevent regurgitation of gastric content during anesthesia induction), on the reported difficulty of laryngoscopy, classified as difficult or not. Due to the fact that there have been no randomised trials, they used an observational database. The problem of using these data to decide whether applying cricoid pressure makes it easier to carry out a laryngoscopy is that factors that determine whether cricoid pressure is used or not may also determine whether the laryngoscopy is difficult, in other words the relationship may be confounded by other factors. The conventional methods of allowing for these confounding factors are linear or logistic models. However, increasingly investigators have used the technique of propensity scores. These have certain advantages, and some disadvantages, over conventional modelling. The method was first described by Rosenbaum and Rubin[2] in 1983 and since has acquired a central place in observational research, being used in many settings. Useful, relatively non-technical, discussions of the use of propensity scores have been given recently [3-5] and which form the basis of this review.

The basic idea

In an observational study, one may wish to study if a particular treatment is associated with the outcome (eg cricoid pressure and mortality). In order for this result to be valid, you need to be sure that there weren't any factors that made it more likely one group would get the treatment and which also would have affected mortality (i.e. confounders.) The essential idea of propensity scores is to use covariates to predict whether a patient receives the intervention (eg application of cricoid pressure) or not. The probability of getting the intervention is the propensity score. The basic premise of Rosenbaum and Rubin[2] is that the score is made up of all possible predictors of the outcome; there are no unknown confounders. This is sometimes known as the ignorability assumption, since it says we can safely ignore other covariates. We will consider whether this is a safe assumption later, but it is often overlooked by propensity score enthusiasts.

Now, consider two patients with the same propensity score; one has cricoid pressure applied and the other has not. Whatever the score, if we assume there are no unmeasured confounders and that these two are the only ones to have this score, then there is a 50% chance that a particular individual of the pair has cricoid pressure is applied. In this way, propensity scores can be likened to a RCT where pairs are matched and then randomised to treatment or not. We can now compare outcomes for these two patients, (eg using a simple chi-squared test in this case since the outcome is binary) safe in the knowledge that the confounders are now balanced. This method is known as matching and is essentially the method used by Caruana et al [1]. However, it is rare that patients will have exactly the same score, and inexact matching causes bias. One option includes discarding patients with no matches, but this can lead to loss of statistical power, so a variety of methods using closest matching criteria are used. There are three other ways of controlling for the propensity score: stratification of the propensity score; weighting by the inverse of the propensity score and including the propensity score as a covariate in a further model. Each method has its plusses and minuses and in general the one least frequently used is the last one, since this requires further assumptions for validity. It is worth noting the balancing aspect of propensity scores is a large-sample property. Thus, a large sample is needed at each value of the propensity score to achieve balance. Exactly how to 'match' is a contentious issue and a recent paper, suggests that matching, in general should not be used for propensity score analysis. [6]

Another question is which variables to choose to form the propensity score. The perceived wisdom was to include as many as possible, but this has led to concerns of overfitting [3]. The propensity score model should include all true confounders to avoid bias. Caruana et al [1] use three different methods for choosing the score and compare them. The authors distinguish between three different types of covariate: Confounders which are related to both treatment (termed exposure in epidemiology) and outcome; prognostic variables which are related to outcome but not treatment, and instrumental variables which are related to treatment but not outcome. A few comments on these definitions are warranted. Epidemiologists would usually add the caveat that a confounder should not be on the causal pathway. The definition of prognostic variables usually includes confounders and so we would not usually exclude confounders from the definition of prognostic variables as do Caruana et al..

Since statistical significance is a poor indicator of whether a variable will induce balance (Caruana et al's model 3), and the method of including all possible variables possibly overfits the data (Caruana et al's model 1) , it is no surprise that model 2 in the paper, which is a judicious balance of included variables, performs better than either models 1 or 3. Further discussion is given in Box 1.

Box 1 Further discussion of the variable types presented by Caruana et al.

Gender would appear to be a prognostic factor in their study, but not a confounder, since gender is predictive of outcome but the proportion of females who got cricoid pressure is similar to those who did not. It is interesting that of the three variables selected as prognostic (gender, patient position and obesity) using multivariate logistic regression, patient position differs between the two treatment groups and so is also a confounder. Instrumental variables are strongly associated with treatment, but not outcome. For example if some individuals always applied cricoid pressure and others never, then a 'natural experiment' would be to compare the outcome between patients treated by these two groups of individuals. However, we would get biased estimates if there was an association between certain individuals and outcome, which did not occur simply because some individuals were more likely to use one treatment over another. Altered neurological status would appear to be an instrumental variable in that there is a big difference in the proportion with and without cricoid pressure treatment, but apparently it is not prognostic of outcome. With regard to propensity scores, Brookhart et al [7] show by simulation that if a variable is related to treatment but not outcome in an observational study, then any bias in estimate of the treatment effect is not reduced, but its estimated standard error may increase, suggesting it is not a good idea to include them in a propensity score which is presumably why Caruana et al excluded them.

Diagnostics

In a randomised trial a 'successful' randomisation will balance known predictors. Similarly we would like to see to see balance in the propensity score. Carnua et al [1] used the difference in mean propensity scores for those treated and not treated, divided by the standard deviation (the standardised mean difference or SMD) and also a new method the weighted balance measure (WBM) which takes into account the strength of the covariate and outcome. This was used because there is less bias if a covariate weakly prognostic of outcome is unbalanced than if one that is strongly prognostic of outcome is unbalanced. A graph showing the distribution of the propensity score by treatment group is a useful adjunct to these quantitative methods, since lack of overlap is immediately apparent, and also a graph of the treatment effect by grouped levels of the propensity score shows whether the treatment effect is constant [4].

Propensity scores versus conventional regression modelling

The propensity score literature is attempting to bridge the two different worlds of epidemiology and clinical trials, and sometimes different terminology is used. For example epidemiologists may use propensity score to assess effects such as smoking or the environment as well a treatment and so refer to 'exposure' whereas a triallist would refer to 'treatment'; for evaluating an intervention such as a drug they mean to the same thing. An epidemiologist will be very cautious about the term 'causal', and will invoke criteria such as those of Bradford-Hill to bolster their claims [8]. In trials where randomisation occurs, there are fewer inhibitions in this sphere, but it is perhaps unfortunate that some users of propensity score methods, which are after all based on observational data, take on the confident tone of triallist with regards causality. A discussion of the results of the analysis of longitudinal data of an intervention with comparable results of clinical trials in the same intervention are given in Box 2.

Traditionally epidemiologists would fit linear or logistic regression models to observational data to examine treatment effects. They would include treatment as a covariate in the model, alongside potential confounders [9].

However, propensity score methods are often seen as more robust to model misspecification than conventional regression models [3]. Conventional models are essentially 'linear'. For example, the effect of an increase of one year of age on the outcome may be assumed the same whether the subject is 20 or 80, and also whether the subject is male or female. The assumption is particularly important if, say, most of the young people get one treatment and most of the old get another, since one then has to extrapolate the figures in order to compare treatment effects in patients of the same age. Incorrect modelling is less important for propensity scoring as long as balance is achieved[3]. It is important to recall, however, that balance of observed covariates does not guarantee balance of unobserved covariates. Omitting a confounder from the propensity score model produces biases similar to those produced by omitting a confounder from a conventional regression model.

It is known that the standard error of the treatment effect estimated from a propensity score analysis will be larger than that from a correctly specified conventional regression model. However, if the conventional regression model is incorrectly specified, the exposure effect estimate will be biased.

When the within-strata treatment effects differ, conventional regression offers the possibility of investigating interaction terms. A disadvantage of the propensity score is that it is a 'black-box'. Conventional regression modelling enables one to see how known confounders perform in the current study compared with earlier studies and in general there is greater familiarity with testing model structures. Thus for example one could compare the effects of known confounders with those found by other investigators, to gain reassurance that the dataset is not unusual.

In general, there are likely to be advantages to both conventional regression modelling and propensity score methods in most situations.

Conclusion

It should be recalled that no observational study can give conclusive proof of causality, and propensity scores are based on observational data. However, randomised trials also have problems and in many cases may be impossible to conduct. A well conducted propensity score study, with careful consideration of possible unmeasured confounders, and with checks which would include the overlap of the propensity score between treated and controls and the relation between the treatment effect and the propensity score, is likely to give a good and precise estimate of a true treatment effect.

References

[1] Caruana E, Chevret S, Pirracchio R Effect of cricoid pressure on laryngeal view during prehospital tracheal intubation: A propensity based analysis. Emergency Medicine Journal 2016

[2] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41-55.

[3] Williamson W, Morley R, Luca A, Carpenter J. Propensity scores: From naïve enthusiasm to intuitive understanding. Statistical methods in Medical Research 2012;**21**:273-93

[4] Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I . Making inferences on treatment effects from real world data: propensity scores, confounding by indication and other perils for the unwary in observational research . BMJ 2013 ; 347: f6409

[5] Austin P. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 2011;**46**, 399-424,

[6] King G and Nielsen R . Working Paper. "Why Propensity Scores Should Not Be Used for Matching". http://j.mp/1FQhySn Last accessed 16 December, 2016.

[7] Brookhart MA1, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.

[8] https://en.wikipedia.org/wiki/Bradford_Hill_criteria Last accessed 16 December, 2016.

[9] Campbell MJ. Statistics at Square Two. 2nd ed Oxford BMJ Books: Wiley, 2006

[10] Shah BR, Laupacis A, Hux JE and Austin PC. Propensity score methods gave similar results to traditional regression modelling in observational studies: a systematic review. J Clin Epidemiol 2005; 58: 550‑559.

[11] Hemkens LG, Contopoulos_Ioannis DG, Ionannidid JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomised trials: meta-epidemiological survey. BMJ 2016; **352**:i493

Box 2 Comparison of the results of propensity score modelling in observational studies with those of clinical trials

A number of investigators have compared the results of analysis of observational data using propensity scores with that of a conventional trial of the same intervention. Shah et al [10] used both regression adjustment and propensity score methods to estimate treatment effects. Although similar effect sizes were reported, estimates obtained using propensity score methods tended to be modestly closer to the null compared with when regression-based approaches were used for estimating odds ratios or hazard ratios.

However, recently Fremantle et al {4} showed that the hazard ratio from an observational data base analysis of the drug spironolactone for treating severe heart failure to reduce mortality suggested the drug was harmful, whereas several randomised trials have shown it to be beneficial and the hazard ratio for the two methods differed by 6.4 standard errors! Freemantle et al showed that the treatment effect was greatest for low propensity scores and suggested that the prescriber making the clinical decision to treat used additional important information on severity of heart failure that the propensity score did not capture, and so the match was made with inappropriately low risk individuals, i.e the decision to prescribe was not an ignorable confounder. In contrast, Hemkens et al [11] looked at 16 propensity score studies and 36 subsequent published randomized controlled trials investigating the same clinical questions (with death as an outcome). Trials were published a median of three years after the corresponding propensity score study. For five (31%) of the 16 clinical questions, the direction of treatment effects differed between the propensity score study and the trial. Confidence intervals in nine (56%)propensity score studies did not include the RCT effect estimate. Overall, propensity score studies showed significantly more favourable mortality estimates by 31% than subsequent trials (summary relative odds ratio 1.31 (95% confidence interval 1.03 to 1.65; I2=0%)). They concluded that studies of routinely collected health data could give different answers from subsequent randomized controlled trials on the same clinical questions, and may substantially overestimate treatment effects. Although they do not comment, presumably there were unmeasured confounders in studies where results differed. They advised that caution is needed to prevent misguided clinical decision-making.